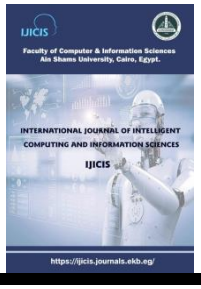
IJICIS, Vol.25, No.1, 51-73
DOI: 10.21608/ijicis.2025.375039.1388

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



GENETIC BIOMARKERS DETECTION FOR ALZHEIMER'S DISEASE

Reham A. Shafik*

Information Systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt.
reham_ashraf@cis.asu.edu.eg

Mahmoud Mounir

Information Systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt.
mahmoud.mounir@cis.asu.edu.eg

Yasmine M. Afify

Information Systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt.
yasmine.afify@cis.asu.edu.eg

Nagwa Badr

Information Systems,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt.
nagwabadr@cis.asu.edu.eg

Received 2025-04-12; Revised 2025-04-12; Accepted 2025-04-13

Abstract: *Alzheimer's disease remains a complex condition with an unclear cause and no known cure. Current treatments focus on symptom management and slowing disease progression. Research is ongoing to uncover its underlying mechanisms, develop effective treatments, and explore early detection and prevention strategies. Genetic data plays a crucial role in Alzheimer's detection, offering significant advantages. Genome-wide association studies (GWAS) have identified numerous genetic variants linked to the disease. Large-scale genetic analyses help researchers understand disease pathways, identify potential drug targets, and contribute to novel therapeutic developments. This review aims to highlight research gaps and limitations while proposing future directions for advancing the field. It provides a detailed survey outlining essential criteria for improving genetic-based detection methods. Researchers can enhance accuracy by selecting optimal approaches for genetic analysis. The review focuses on recent studies that integrate genetic data with artificial intelligence (AI) to identify mutated genes associated with Alzheimer's and classify the disease efficiently. Findings indicate that, despite a relatively small body of published research, studies in this field have grown exponentially since 2020. This review offers a comprehensive analysis of genetic and AI-driven approaches for Alzheimer's detection. It serves as a valuable resource for researchers, clinicians, and policymakers, shedding light on the current state of the field, guiding future research, and supporting the development of more accurate and effective early detection methods for Alzheimer's disease.*

Keywords*: Alzheimer's detection, Artificial Intelligence, Genetic datasets, Gene Biomarker.*

***Corresponding Author**: Reham A. Shafik

Information Systems Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: reham_ashraf@cis.asu.edu.eg

1. Introduction

Alzheimer's disease is considered a progressive condition characterized by neurodegeneration, mainly affecting the brain, causing cognitive deterioration, loss of memory, and alterations in behavior and personality. It is the primary cause of dementia, accounting for roughly 60–80% of all cases [1]. As the disease progresses, individuals may experience forgetfulness and mental disorientation, language and communication difficulties, impaired judgment, disorientation, and changes in mood and personality, and eventually a loss of the ability to perform daily activities. The disease is typically considered a complex disorder influenced by a combination of genetic, environmental, and lifestyle factors. There are specific genetic mutations associated with Alzheimer's disease. A key area of research involves identifying complex diseases linked to genetic variations within the human genome. Genome-Wide Association Studies (GWAS) focus on detecting genetic variants, particularly those associated with complex diseases through single nucleotide polymorphisms (SNPs). In recent years, artificial intelligence (AI) techniques have played a crucial role in genetic disease detection, offering the capability to process large-scale genetic data, uncover significant patterns, and enhance prediction accuracy [2]. AI approaches have the potential to revolutionize genetic disease detection by leveraging the power of data analysis, pattern recognition, and prediction. By enhancing our ability to analyze and interpret genetic data, AI approaches can effectively enhance preliminary screening for Alzheimer's disease by leveraging advanced data analysis techniques, identifying biomarkers, and providing predictive models. Early detection enables timely interventions, improves patient outcomes, and facilitates research efforts to find more effective treatments and preventive strategies for this devastating disease [3].

This paper discusses the recently published approaches for preliminary screening for Alzheimer's disease (AD) by the detection of genetic variants causing the disease, including the datasets, AI algorithms and their results comparisons from the last 4 years (2020 - 2023). Aim includes summarizing the studies of selected published research papers during this period and mentioning the challenges that face the research in this direction and how they could be handled.

What makes this paper different is the concentration on the detection approaches that depend on genetic data for the aim of improving research in this direction. Genetic data plays a crucial role in advancing Alzheimer's disease research. By studying the genetic factors associated with the disease, researchers can identify potential therapeutic targets, develop animal models for studying the disease, and improve the overall understanding of its complexity. This knowledge can pave the way for new diagnostic methods and more effective future treatments.

This paper provides answers to several research questions that may arise when working on Alzheimer's detection using genetic data, including: 1) why use Genetic data? 2) What are the most applied methods & approaches? 3) What are the most common data sources and datasets? 4) What are the challenges that researchers face in this field & how they can be handled?

The paper structure is as follows. Section 2 shows the methodology of the study, inclusion, and exclusion criteria. Section 3 presents a summary of the literature review during the assigned period. Section 4 compares

and analyzes the data used and methods. Section 5 presents challenges of using Genetic data in disease detection and how they can be handled. Finally, the review discussion and conclusion in section 6.

2. Methodology

The methodology used in this paper is explained in this section. The criteria of selection, the research progress, and results are presented in the following subsections.

2.1. Criteria of Selection (Inclusion/Exclusion)

The following inclusion criteria were used to find relevant studies: 1) any publication, except systematic reviews, that discusses the detection of AD using genetic data, detecting the gene biomarker that addresses its presence or the risk of its development; 2) publications that combine the genetic data with the image data to improve the detection accuracy. 3) published between January 2020 and December 2023; 4) papers from journals, conferences, or workshops; 5) written in English with no regard to location; 6) not duplicated.

Due to the direction of this research, it is important to mention that any papers using only imaging data; not considering the disease genetic variations; are excluded.

2.2. Research Procedure

As previously mentioned, the focus of this research is considering the genetic variants that cause the disease while developing the approach of its detection. The key term "AD genetic biomarker detection" is used throughout Science Direct (https://www.sciencedirect.com), Pumbed (https://pubmed.ncbi.nlm.nih.gov), Frontiers (https://www.frontiersin.org), PLoS ONE (https://journals.plos.org/plosone), MDPI (https://www.mdpi.com), IEEE explore (https://ieeexplore.ieee.org), Elsevier (https://www.elsevier.com), and Springer nature (https://link.springer.com) databases. These databases were selected based on their well-established credibility or prominence.

All titles and abstracts of the publications were reviewed to check if they match the inclusion requirements. Subsequently, a thorough re-evaluation of all research papers was conducted, specifically examining the body and conclusion sections to ensure their alignment with the criteria of this review. With the research questions in focus, a comprehensive investigation of the relevant papers was carried out to extract the essential information from the eligible studies, that are: 1) The publication year; 2) The paper objective including gene biomarker detection, feature selection, develop or a new detection approach; 3) Algorithms & methods used; 4) Used AI: machine learning, deep learning or data mining; 5) Datasets used; 6) Data sources such as Alzheimer's Disease Neuroimaging Initiative (ADNI) and Whole-genome sequencing (WGS); 7) Evaluation measure such as: Recall, F1-Score, Accuracy, Precision, or AUC; 8) Methods validation tools & techniques.

2.3. Results

As a result of the previous procedure, many research papers were excluded as they are using other biomarkers rather than genetic ones. A total of 22 research papers were downloaded and included in this survey. The

distribution of the included papers through the mentioned databases is shown in Fig. 1. It can be concluded that the Elsevier and IEEE Xplore databases have a noticeably higher number of papers included that are related to the research keywords.

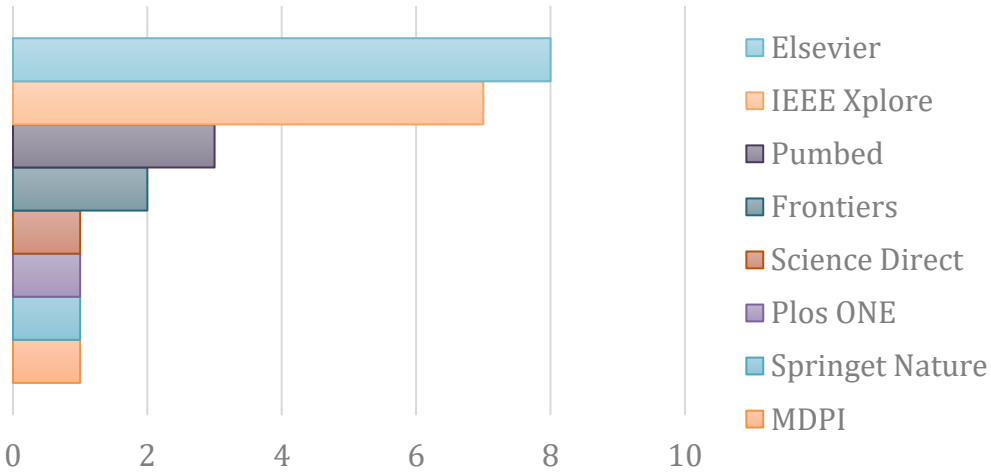


Figure 1 Distribution of included papers across publishers

3. Literature Review

For Alzheimer's detection using AI approaches, variant methods, dependencies, and objectives are taken as a case of study to achieve the most optimal accuracy of classification for the disease. In this section, some of these cases published throughout the last 4 years are presented.

It is noted that the most common objective through all is to improve the accuracy of classifying the presence or absence of the disease through finding the most related genes to it, or using new improved models for preprocessing, extracting features, and classification approaches.

In [4], a method was proposed to identify biomarker genes that may serve as risk factors, using gene expression information derived from four microarray datasets to investigate Alzheimer's disease (AD). Various feature selection methods and classification models were employed to determine the most informative genes and assess their performance. Among the classifiers, Support Vector Machine (SVM) with a linear kernel achieved the highest accuracy, with 93.8% on training and 89.8% on testing. In contrast, the C4.5 decision tree and Naïve Bayes classifiers recorded the lowest accuracy. SVM with a linear kernel was further utilized to assess the quality of selected genes using different feature selection methods. The study concluded that incorporating Principal Component Analysis (PCA) in feature selection helped eliminate highly correlated genes. However, since the selection techniques were applied to a small and specific dataset, the findings may lack generalizability and accuracy when applied to larger datasets.

The study in [5] focused on utilizing machine learning (ML) techniques to discover biomarkers associated with Alzheimer's disease (AD). Several learning algorithms, like Naïve Bayes (NB), Logistic Regression

(LR), Random Forest (RF), and SVM, were applied to genetic data from the AD Neuroimaging Initiative Phase 1 (ADNI-1) and Whole Genome Sequencing (WGS) datasets. In the whole-genome ADNI-1 approach, the classification models achieved overall accuracies of 98.1% (NB), 97.97% (RF), 95.88% (SVM), and 83% (LR). The findings demonstrated that machine learning-based classification techniques show promise for the early detection of AD.

The study in [6] investigated the application of a deep neural network (DNN) to predict Alzheimer's disease (AD) by integrating gene expression and DNA methylation data. The primary aim was to build a highly accurate and reliable model for early AD diagnosis. The methodology included several steps such as data preprocessing and feature selection across both molecular datasets. Results suggested that combining gene expression and DNA methylation with a DNN could be a promising strategy for early AD detection. However, the study only integrated two molecular layers without using multi-omics data from the same sample cohort. To compensate, the authors synthesized every conceivable sample pair corresponding to each class label, forming a new dataset that may have introduced the risk of model overfitting.

Other work by [7] aimed to predict the impact of genetic alterations linked to Alzheimer's disease. They developed a predictive model that can assess the functional consequences of these variants, which can help clarify the genetic mechanisms involved in the disease and potentially guide future therapeutic interventions. They employed various machine learning algorithms and feature engineering techniques to train a predictive model. The model aimed to classify the variants into different categories based on their potential functional impact. The approach achieved an accuracy of 81.21% during 10-fold cross-validation and 70.63% on an independent test set comprising 5,785 variants. They were having limitations with the server used, it can only process 1,000 variants at a time due to its high computational cost.

The study in [8] involved collecting genetic data from individuals with Alzheimer's disease and healthy controls to develop a classification model using various machine learning algorithms. The goal was to differentiate between AD patients and non-AD individuals based on their genetic profiles. A two-step feature selection process was applied, resulting in a refined list of key genes. Among the tested classifiers, XGBoost performed best when trained solely on gene expression features, achieving an Area Under the Curve (AUC) of 0.64. When additional features such as age and years of education were incorporated, the AUC slightly improved to 0.65. However, these results suggest that higher accuracy may still be attainable.

The study in [9] introduced a novel approach to identifying hub genes associated with Alzheimer's disease (AD), as these genes exhibit high connectivity within a gene network and may play crucial roles in biological processes. Microarray gene expression datasets from six brain regions, available publicly, were analyzed using two distinct approaches. The initial approach sought to pinpoint biomarker genes in each region and identify the region that provided the best efficacy for Alzheimer's disease (AD) detection. The second approach involved constructing a gene network by selecting the key gene in each region and identifying hub genes, which were then used in a classification model to assess their ability to distinguish AD patients from controls. The first approach achieved 95.7% accuracy in the Entorhinal Cortex (EC) region, while the second approach demonstrated that hub genes improved classification accuracy, reaching 100% for the EC region. However, the study acknowledged that relying solely on brain gene expression data might limit the generalizability of the model's findings.

In [10], the authors created a gene selection process that combines filter, wrapper, and unsupervised techniques to identify pertinent genes. Their approach combines Wrapper-based Particle Swarm Optimization (WPSO), Minimum Redundancy Maximum Relevance (mRmR), and an Autoencoder to extract Alzheimer's disease (AD)-related features. After selecting the relevant genes, they developed an enhanced Deep Belief Network (IDBN) with a basic termination criterion. To optimize the IDBN's hyperparameters, they employed Bayesian Optimization. The developed gene selection process (mRmR-WPSO-AE), combined with IDBN, achieved an impressive classification accuracy of 96.78% for Alzheimer's.

In [11], a deep learning-based classifier, coupled with an embedded feature selection approach, was employed to categorize Alzheimer's disease (AD) patients using DNA methylation data. The data underwent preprocessing, including quality control, normalization, and downstream analysis, before selecting relevant features. To address the large number of associated CpG sites, four embedded feature selection models were contrasted. The results indicated that the tree-based embedded approach, AdaBoost, achieved higher accuracy than the other three methods and was therefore chosen for the targeted classification approach. Subsequently, an Enhanced Deep Recurrent Neural Network (EDRNN) was developed and assessed in comparison to existing classification models, such as a Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Deep Recurrent Neural Network (DRNN). The findings demonstrated that the EDRNN outperformed the other techniques.

The study in [12] proposed a three-step deep learning method (SWAT-CNN) for identifying genetic variants linked to disease phenotypes, particularly focusing on SNPs. This method was designed to improve the accuracy of disease classification models by pinpointing phenotype-related SNPs. The approach successfully identified the APOE region as the most significant genetic locus for Alzheimer's disease (AD), and the classification model achieved an area under the curve (AUC) of 0.82.

In [13], researchers developed a machine learning approach to predict Alzheimer's disease (AD) using DNA methylation and gene expression datasets. To tackle the challenges associated with high dimensionality and small sample size (HDLSS), they utilized an autoencoder (AE) to create a compact and continuous feature representation. Various machine learning models were then applied to classify AD based on the encoded features, with performance evaluated using accuracy and area under the curve (AUC) metrics. Their findings demonstrated that integrating DNA methylation and gene expression data enhances AD indication accuracy. The proposed method surpassed the current state-of-the-art approach, improving accuracy by 9.5% and AUC by 10.6%.

The study in [14] aimed to predict and diagnose AD during its initial stages with high classification accuracy using SNP biomarkers. To address the challenge of high-dimensional feature space, the researchers developed a thorough framework for the early detection of AD. and identification of key genes through SNP analysis. They applied both filter and wrapper methods to assess feature relevance based on correlation with dependent variables and the impact of selected feature subsets in model training. Two feature selection techniques were employed to identify the most impactful AD-related genes. Gradient Boosting Tree (GBT) was implemented on genetic data from the Alzheimer's Disease Neuroimaging Initiative Phase 1 (ADNI-1)

and whole-genome sequencing (WGS) datasets. The results indicated that the GBT model achieved an accuracy of 99.06% using Boruta for feature selection, while applying information gain resulted in an average accuracy of 94.87%.

In [15] a simulative deep learning model was implemented by the using of chromosome 19 genetic data. By employing the occlusion method, the model assessed the influence of individual SNPs and their combined effects on the probability of developing AD. Specifically, it identified the top 35 SNPs with AD risk located on chromosome 19 and examined their capability to predict the rate at which AD advances.

The objective in [16] was to create a paragenic risk score (PRS) that goes beyond utilizing single genetic marker data. It includes features related to epistatic interactions and employs machine learning methods to predict the lifetime risk of late-onset Alzheimer's disease (LOAD). The approach introduces two innovations compared with typical PRS models, that are, the direct integration of epistatic interactions between SNP loci through an evolutionary algorithm that considers shared pathway information, and, instead of relying solely on simple logistic regression, it estimates the risk using an ensemble of machine learning models, specifically deep learning and gradient boosting machines. The results obtained an AUC of 83%.

The work in [17] aimed to identify specific SNPs that represent groups of genetic variations across the entire genome. These SNPs were selected to create a generalized PRS model that can be applied across different cohorts and genotyping panels. The PRS modeling was conducted on five separate cohorts focused on AD development. The most effective models were those that identified shared groups of genetic variations contributing to the differentiation between AD cases and controls across multiple cohorts. These identified groups were utilized to create a generalized PRS model, which was then tested in the initial five development cohorts as well as three additional AD cohorts. The developed model achieved an average discriminability accuracy of over 70% across multiple AD cohorts. It incorporated variants from several widely recognized genes associated with AD risk.

The aim of [18] was to detect SNP biomarkers associated with AD for reliable disease classification. Deep transfer learning is employed with various experimental analyses to achieve a dependable classification of AD. They initially trained CNN on the genome-wide association studies (GWAS) dataset generated from the AD neuroimaging initiative. Deep transfer learning is then applied to further train the CNN base model on a separate AD GWAS dataset, enabling the extraction of the final feature set. The extracted features are then utilized for AD classification through SVM. The model achieved an accuracy of 89%.

In [19], the GWAS dataset, which includes key genetic markers for complex diseases, was analyzed. The dataset contained 620,901 attributes, requiring a complex feature selection approach to refine the most relevant predictors of Alzheimer's disease (AD). This method combined association tests, PCA, and the Boruta algorithm to detect significant features. The selected features were then used as input for wide and deep neural network models to classify AD cases and healthy controls. Results demonstrated high performance, achieving 99% accuracy and F1-score. However, the model faced limitations because of the relatively small sample size and the fact that the original dataset contained more features than samples.

The study in [20] aimed to find the most effective model for identifying biomarker genes related with AD by evaluating various feature selection methods. A gene selection approach was developed, integrating filter, wrapper, and unsupervised techniques to identify the most related genes. The performance of multiple feature selection methods, including mRMR, Chi-Square Test, CFS, F-score and GA, was assessed using an SVM classifier, with accuracy measured through 10-fold cross-validation. These methods were applied to a benchmark AD gene expression dataset containing 696 samples and 200 genes. The results indicated that mRMR and F-score, combined with an SVM classifier, effectively identified biomarker genes, achieving an accuracy of 84%.

The study in [21] focused on analyzing inflammatory gene expression patterns in the parietal cortex (PCx) and temporal cortex (TCx) using a human brain RNA sequencing dataset. The goal was to uncover genetic links associated with dementia and improve its prediction. By applying five machine learning and statistical methods, the study demonstrated that gene expression patterns in the PCx provided better detection and classification of dementia patients compared to those in the TCx. The results provide evidence that the PCx contributes significantly to neuroinflammatory processes linked to dementia. Among the applied models, logistic regression (LR) achieved the highest classification accuracy with 77.54%.

The study in [22] introduced Hygieia, an AI/ML-ready pipeline designed to integrate genomics and clinical data for disease prediction and gene association analysis. Hygieia supports datasets of varying sizes and granularity, enabling accurate identification of genes linked to specific disorders. It employs supervised learning method to examine combined gene expressions and multivariate clinical data. Additionally, the pipeline incorporates a Random Forest-based model for regression analysis and disease prediction, operating with no need for hyperparameter tuning.

This study [23] focused on detecting the most impactful SNPs by analyzing gene-gene interactions. The researchers applied their framework to the ADNI dataset, utilizing ensemble learning and the MDR constructive induction algorithm to efficiently uncover epistatic interactions related to Alzheimer's disease (AD). The classification accuracy of five-way interaction models ranged from 0.8674 to 0.8758, while two-way, three-way, and four-way models achieved accuracies between 0.6515–0.6649 and 0.7071–0.7170, respectively. Furthermore, 76 SNPs were mapped to 38 genes, including both previously recognized AD-related genes and newly discovered ones.

The target in [24] was to develop a prediction model for LOAD subtypes. A deep neural network (DNN) architecture comprising six hidden layers with 512 neurons each was employed, utilizing RReLU activation, a 50% dropout rate, and a batch size of 32 to predict principal component (PC) scores from variant data in the discovery cohort. The networks were trained for 100 to 1500 epochs at intervals of 100. The best-performing model reached an accuracy of 0.694. Additionally, a DNN was used to simultaneously predict LOAD subtypes and associated phenotypes, demonstrating high accuracy in both the discovery and validation cohorts.

Authors of [25] targeted the identifying of potential genetic epistasis related to AD using the linear regression method. They performed genome-wide SNP-SNP interaction on cerebrospinal fluid Aβ42, using age, gender,

and diagnosis as covariates. Results replicated 100 AD-related genes that are previously recorded, and 5 gene-gene interaction pairs were found to be overlapped with the PPI network.

In [26], an artificial intelligence-driven approach integrating both unsupervised and supervised machine learning was used to detect novel gene candidates for further investigation. The study uncovered biologically relevant and functionally connected genes associated with AD. This approach involved a meta-analysis of microarray datasets derived from the frontal cortex and cerebellum of individuals with AD. Functional network analysis revealed two downregulated genes, ATP5L and ATP5H, both encoding subunits of ATP synthase (mitochondrial complex V), which may play a role in AD progression.

The review of studies on AD detection indicates that numerous existing models have achieved high levels of accuracy using specific approaches. These promising results have to inspire researchers to apply these approaches to diverse datasets, aiming to predict novel genetic features of the disease and improve the accurate classification of its presence in early stages. Additionally, certain studies may recommend suitable datasets for testing with more advanced models, serving the same purpose. Each applied approach has the potential to yield better outcomes or identify new genes associated with the disease, which could significantly contribute to faster disease detection.

4. Comparative Analysis

To develop or apply new effective approaches, researchers should start their investigation from the best last reached point. A comprehensive analysis of the included papers determines that there is a common workflow for this process, which is presented in Fig. 2. Commonly the process starts with selecting the best dataset to work on depending on the main aim of the research, then obtaining the data access through contacting the provider or what else needed and downloading the dataset, start understanding the structure of the downloaded data by reading the documentation attached to it. The next step is the preprocessing of the data, preprocessing step in genetic data analysis is of utmost importance as it plays a crucial role in ensuring the accuracy, reliability, and meaningful interpretation of the data. Genetic data preprocessing involves a series of steps aimed at cleaning, transforming, and preparing the raw genetic data for downstream analysis, the commonly used preprocessing was the data cleaning, quality control, data normalization then linking the genetic data with clinical or phenotype information to get more accurate classification of the disease, also feature reduction is used with large datasets with a high number of features, to help reduce the dimensionality of the data, resulting in faster and more efficient analysis, by eliminating irrelevant or redundant features. The third step is the feature selection, calculating PRS providing a quantitative measure of an individual's genetic risk for the disease trait, and then the ML selection approaches are applied with validation step to assess the performance and generalization of the selected features. The fourth step is the classification of selected features, it usually requires trying different algorithms and testing their results, to get the best algorithm that fits with the selected type of data. Finally, Cross Validation techniques are applied or performance comparisons so the best results can be reached.

The results of the different datasets, feature selection, classification, and validation methods that were previously applied are discussed in the following subsections.

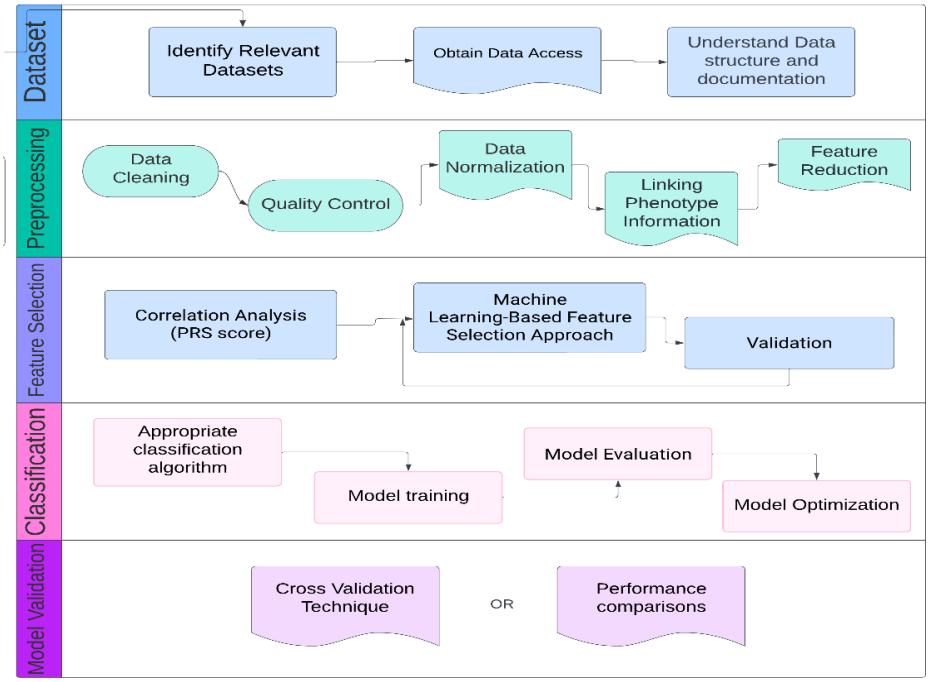


Figure 2 The common process workflow.

4.1. Sources & Datasets

AD genetic data, especially large-scale datasets with comprehensive genetic information, can be limited in availability. Obtaining access to such datasets may require collaborations with research institutions, adherence to data sharing agreements, or participation in specific research consortia. In this section, we are going to discuss the frequently used data sources and datasets throughout the review period.

4.1.1. Data sources

Fortunately, there are several sources where AD genetic datasets can be accessed, some of these sources are available online and require a quick simple registration for data access. Recent studies use this available source of data which added their value with much informative data. Fig. 3 presents the use of some of the available sources of data in research through recent years. The increase of the use of some sources can be concluded, which addresses their added value in the research field.

Based on Fig. 3, it can be concluded that the use of these repositories increases in recent research especially for the ADNI database which is used by nearly 43% of the published studies in this field throughout the review period.

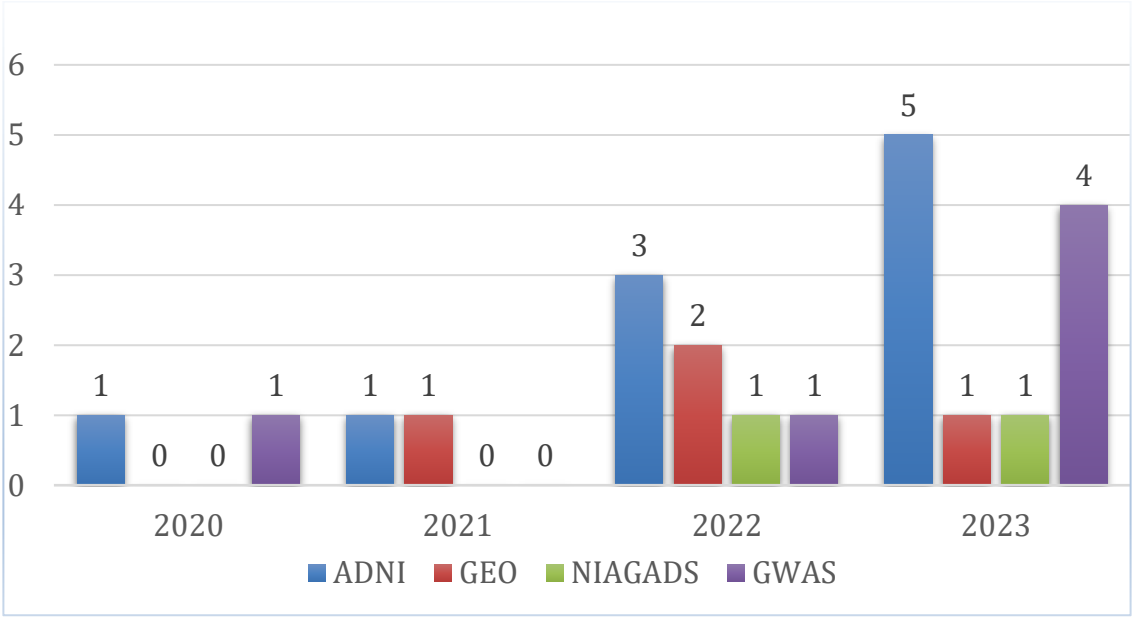


Figure 3 AD Genetic repositories usage through years.

4.1.2. Datasets

The consideration of the dataset is very effective when applying any of the proposed approaches. Some datasets are frequently used in several applied approaches and reach different accuracies. Fig. 4 shows the average accuracy of each dataset while the number of uses of each dataset is shown in Table 1.

DNA methylation data provides valuable insights into the epigenetic regulation of gene expression and its association with various biological and disease processes. It serves as a critical resource for studying epigenetic modifications and their implications for health and disease. It's noticeable that integration with methylation data plays a role in improvement of the AD classification and biomarker detection.

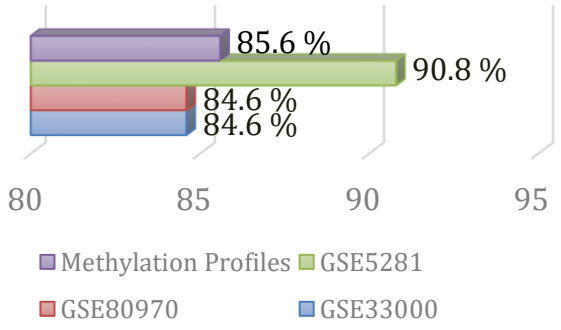


Figure 4 Average Accuracies of the Frequently Used Datasets

DATASET	# USES
GSE5281	3
GSE33000	4
GSE76105	4
Methylation Profiles	3

Table 1 Number of Uses of Datasets

Despite the repeated use of some datasets, some research papers used a novel combination of integrating datasets that achieved high accuracy. Two distinct types of omics datasets—gene expression and DNA methylation profiles—were utilized in [6]. Two large-scale gene expression profiles GSE33000 and GSE44770 were integrated for increasing the sample size, with a specific focus on the prefrontal cortex. Integration resulted in 257 non-demented, and 439 AD samples. In addition, DNA methylation profiles from the same brain region were also incorporated. Also, in [15], Chromosome 19 was used to identify Chromosomal-risk impact score (CRIS) at the individual level attributed to individual SNPs and their interactions with each other by developing a deep learning model. Although, using a single chromosome significantly reduces the computational burden to explore the feasibility and effectiveness of this type of model It was considered that Chromosome 19 is well known to include many AD-linked genes including APOE, APOC1 and TOMM40, that can provide better model results.

4.2. Preprocessing

Genetic data is inherently complex, consisting of a vast number of variables (such as SNPs) and interactions between them. AI algorithms need to effectively handle this complexity and extract meaningful patterns and insights from the data. Also, it can be noisy, incomplete, or contain errors. Ensuring data quality and addressing variability across different datasets (e.g., from different populations or platforms) is crucial for accurate analysis and interpretation. The most applied preprocessing approaches are the Quality Control (Plink analysis), Z-score normalization and PRS calculations as shown in Fig. 5.

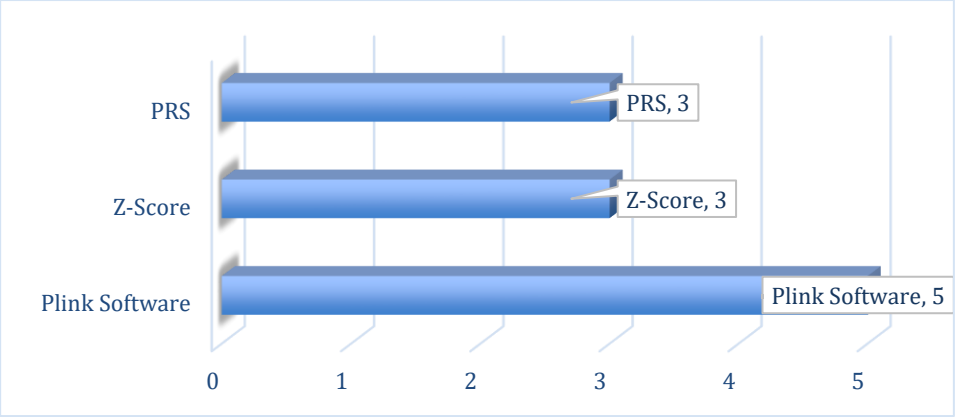


Figure 5 Usage frequency of preprocessing techniques

Quality Control (QC) procedures are performed to ensure the accuracy and dependability of genetic data. This includes checking for sample-related issues such as sex discrepancies, relatedness, and population stratification. Additionally, variant-level QC is performed to evaluate the quality of genetic markers, such as removing low-quality variants with low call rates, high missingness, or deviations from Hardy-Weinberg equilibrium, Plink Software is the most common used for these procedures, PLINK is widely used in genetics research and has been employed in many GWAS and other genetic analyses. Its versatility, extensive functionality, and command-line interface make it a valuable tool for researchers working with genetic data.

PRS are calculated based on the weighted sum of multiple genetic variants associated with AD risk. PRS calculation involves selecting relevant variants, assigning weights based on their effect sizes, and summing them to derive a single score for each individual. It was mainly used for the purpose of detecting disease relevant genes or SNPs.

Normalization of genetic data using z-scores is a common approach to standardize the values of genetic variants in AD research. By normalizing genetic data using z-scores, researchers can remove the influence of scale and variability differences between variants and focus on the relative differences in genotype values across individuals. This normalization method facilitates comparisons, combining data from multiple studies, and identifying potential associations with AD while accounting for inter-individual and inter-variant variations.

In [23], a comprehensive preprocessing approach was employed to integrate additional information about individuals classified as either normal or affected. The process included estimating alleles for SNPs rs7412 and rs429358 to incorporate APOE genotyping. Quality control (QC) procedures were then applied to eliminate SNPs exhibiting poor genotyping performance. To further refine the dataset, SNPs with high correlation were removed using linkage disequilibrium analysis. Statistical association tests were conducted to assess the relationship between each SNP and the disease, applying a p-value threshold of <0.01. Finally, significant SNPs were identified by intersecting the results of these analyses.

4.3. Feature Selection/ Reduction Approaches

GWAS involves high-dimensional data, making direct interpretation challenging, as most SNPs are either not useful or devoid of relevant information. Therefore, identifying the most critical SNPs is essential to enhance the interpretability of machine learning models, reduce model variance and overfitting, and minimize the number of features, ultimately lowering computational costs. At this stage, association analysis results are utilized to extract key features that are strongly associated with the target phenotype. The most used feature selection approaches and their accuracy comparisons are shown in Figs. 6 and 7.

Other unique effective feature selection approaches were applied with different targets. In order to reflect biological processes and examine the interplay between the two types of omics data utilized,[6] suggested approaches based on differentially expressed genes (DEGs) and differentially methylated positions (DMPs). In [11], hybrid feature selection methods were utilized, combining multiple feature selection approaches. Specifically, both filter and wrapper techniques were applied, incorporating two tree-based methods (AdaBoost and Random Forest) along with two regularization-based embedded techniques (LASSO Regression and SVM). In [12] a deep learning-based approach was developed to identify informative SNPs. The entire genome was divided into non-overlapping fragments of optimal size, and deep learning algorithms were employed to select phenotype-associated fragments containing SNPs linked to the phenotype. The optimal deep learning algorithm was then applied using an overlapping Sliding Window Association Test (SWAT) within these selected fragments to calculate Phenotype Influence Scores (PIS), utilizing both SNPs and the phenotype of interest to identify the most informative SNPs. Feature reduction techniques were applied in [23] to address high dimensionality by eliminating redundant and insignificant features from the dataset. The TuRF feature selection algorithm was utilized to enhance the performance of the well-known

ReliefF algorithm. By selecting the top 30% of significant SNPs, the number of SNPs was narrowed down from 3,502 to 1,050. Furthermore, ensemble learning techniques—such as Random Forest (RF) with Gini index and permutation importance, XGBoost, and CART—were utilized to determine the top 20 ranked SNPs identified by each method. In [19], a logistic regression-based association test was conducted to evaluate the relationship between each SNP and Alzheimer's disease (AD). The top 1,000 SNPs were selected based on their significance levels (p-values) and then processed through a hybrid feature selection approach that incorporated PCA and the Boruta algorithm.

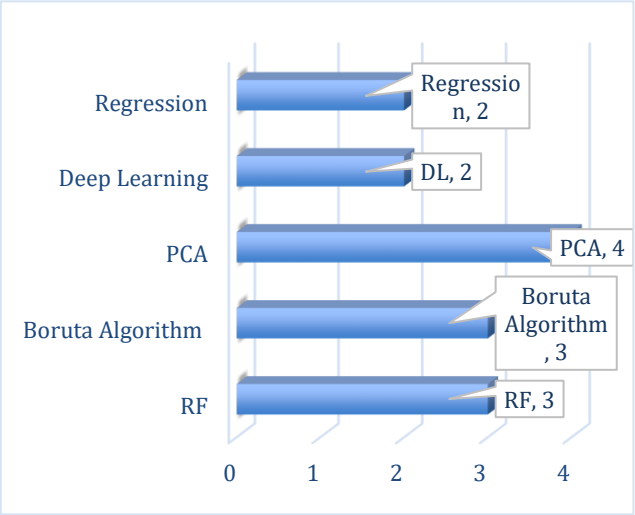


Figure 6 Feature Selection Approaches Frequencies

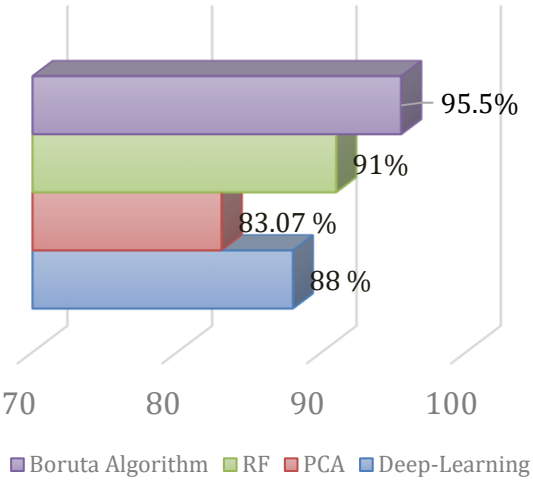


Figure 7 Average Accuracies of the frequently used feature selection methods

4.4. Classification Models

A variety of machine learning algorithms can be used for training the AD prediction model, such as logistic regression, SVM, RF, or deep learning models like neural networks. During training, the model trains the patterns and relationships between the selected features and the disease status by adjusting its internal parameters. Figs. 8 and 9 presents the repetition of frequently applied models for classification during the review period and their average achieved accuracy.

Parameter tuning for Classification and regression trees (CART) using Randomized SearchCV method was applied in [23] to find the best parameters. The CART algorithm assigns importance scores by measuring the reduction in the selected criterion parameter, such as Gini index or entropy, to determine optimal split points. The SNP with the lowest entropy was selected as the best split. The feature importance property was utilized to extract the comparative importance scores assigned to each input SNP. Finally, the CART algorithm identified the top-ranking SNPs to assess their effectiveness in detecting high-ranked epistasis interactions.

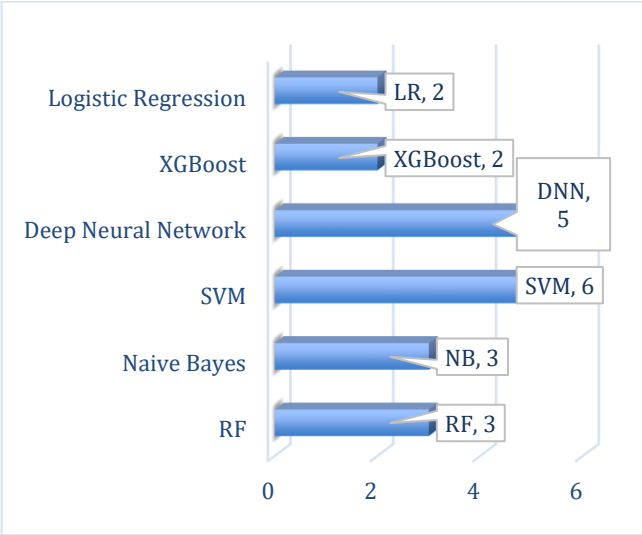


Figure 8 Classification Approaches Use Frequencies

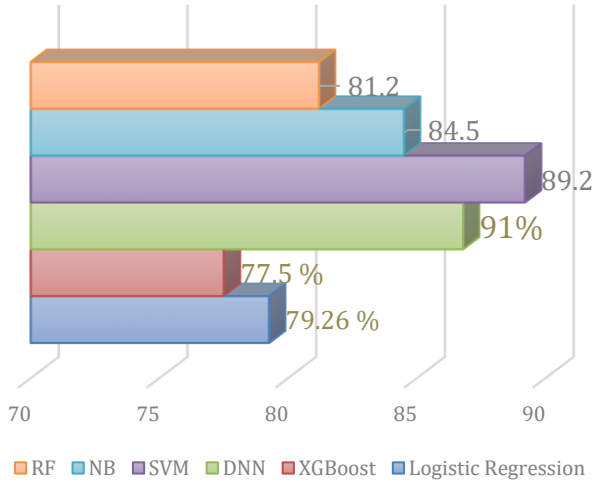


Figure 9 Average Accuracies of the Frequently Used Classification Methods

Another important classifier was developed in [11]. To classify AD cases and controls, an Enhanced Deep Recurrent Neural Network (EDRNN) incorporating stopping criteria was implemented. The model determined the current state based on the previous state and the given set of current input features. Also, WNN was identified as the most effective classifier, with a reliable accuracy rate of 99% in [19]. Different input variations and feature combinations were explored to optimize model selection for effective Alzheimer's disease (AD) classification while identifying the most significant features. The neural network architecture consisted of input, hidden, and output layers, each containing a predefined number of neurons. Two architectural models were employed: a Wide Neural Network (WNN) with one hidden layer containing a large number of neurons, and a Deep Neural Network (DNN) composed of multiple hidden layers, each with fewer neurons.

4.5. Models Comparison / Evaluation Criteria

Evaluation criteria are crucial to assess and compare the performance, applicability, and potential of different AI models in AD detection using genetic data. It aids in selecting the most effective and clinically relevant approaches for early diagnosis, risk assessment, and personalized interventions in Alzheimer's Disease. The Accuracy, Recall, Precision and F1-Score are the most used evaluation criteria in this field, calculated using the below equations.

$$\text{Accuracy} = \boldsymbol{TP + TN} / (\boldsymbol{TP + TN + FP + FN}) \quad (1)$$

$$\text{Precision} = \boldsymbol{TP}/(\boldsymbol{TP + FP}) \quad (2)$$

$$\text{Recall} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FN}) \quad (3)$$

$$F1 - Score = (\mathbf{2} \times \frac{(\boldsymbol{Precision} \times \boldsymbol{Recall})}{\boldsymbol{Precision} + \boldsymbol{Recall}}) \quad (4)$$

5. Discussion

In this section, we elaborate on our insights and findings on using genetic data and AI for Alzheimer's detection research. The following subsections discuss different perspectives to answer the previously mentioned research questions: Q1) Why use Genetic data? Q2) What are the most applied methods & approaches? Q3) The most common data sources and datasets? Q4) What are the challenges that researchers face in this field & how they can be handled?

5.1. Why Use Genetic Data?

AD has a significant genetic component, and certain genetic variants are correlated with an elevated risk of the disease [1]. By incorporating genetic data into AI models, researchers can identify and leverage these genetic risk factors to improve the accuracy of AD detection.

Additionally, Genetic data, when combined with AI algorithms, can aid in the early detection of AD as genetic markers or profiles can provide insights into an individual's predisposition to the disease even before clinical symptoms manifest [2]. Early detection enables timely interventions, potential risk reduction strategies, and personalized treatment plans. AI techniques can harness genetic data to calculate PRS for AD. PRS aggregate data from numerous genetic variants to assess an individual's overall genetic risk for the disease [2]. PRS-based models, when integrated with AI, can enhance prediction accuracy and provide a quantitative measure of AD risk. The integration of genetic data with AI in AD detection paves the way for personalized medicine approaches. By considering an individual's genetic profile, AI models can provide tailored risk assessments, prognosis predictions, and treatment recommendations. This personalized approach improves patient care and aids in advancing precision medicine approaches tailored to AD. By analyzing genetic variants and their associations, AI algorithms can identify potential biological pathways, gene-gene interactions, or molecular processes implicated in the development and progression of AD. These insights contribute to enhancing our understanding of the disease and could facilitate the identification of new therapeutic targets.

5.2. The Most Applied Methods & Approaches and Their Achieved Accuracy

Research in Alzheimer's disease detection using genetic data usually takes one of the following paths: 1) Identify possible genetic biomarkers for AD. 2) Develop or enhance Classification prediction model of the disease to improve the early diagnosis. Tables 2 and 3 address the existing methods and their accuracies for each of the two objectives respectively.

PCA is highly used for feature reduction and classification, but in most cases, it's concluded that it has a negative effect on the approach performance, that might be caused due to complex genetic interactions, high dimensionality and sparsity or population structure and stratification of genetic data [27]. It is important to carefully consider the suitability of PCA for genetic data analysis. Alternative techniques that are specifically designed for genetic data, such as linear discriminant analysis (LDA) or sparse PCA, might be more appropriate in certain cases.

Boruta Algorithm proved its effectiveness as a feature selection approach for both mentioned research objectives, as it was used with different classification approaches (GBT classifier, DNN, NB and SVM) and achieved accuracy that ranges from 95% to 99%.

Table 2. Research summary on detection of new SNPs or gene portions that are signs of the disease presence.

Research	Publication Year	Method Applied	Accuracy
[4]	2020	PCA + Random Forest	0.890
[8]	2021	Logistic Regression (LR)	0.774
[9]	2021	SVM	0.894
[14]	2022	Boruta Algorithm + GBT Classifier	0.949
[12]	2022	deep learning–based approach	0.750
[15]	2023	Deep Learning Model	0.682
[20]	2023	mRMR and F-score filters	0.840
[21]	2023	Multinomial LR with Ridge Estimator, Random Tree (RF)	LR(0.7742),RF(0.6452)
[25]	2023	GPU-based Linear Regression Model	
[26]	2023	K-means clustering + regression trees	0.880
[16]	2023	Baseline model + PRS mode + Epistatic model	0.829
[17]	2023	Enhanced PRS Model	0.70
[23]	2023	Ensemble learning methods	0.800

Table 3. Research summary on feature selection and Classifying the presence or absence of the disease.

Research	Publication Year	Feature Selection	Classification Approach	Accuracy
[5]	2020	Boruta Algorithm	NB, RF, SVM, and LR	0.981, 0.979, 0.958 and 0.83
[6]	2020	DEG	Deep neural network (DNN) with BAYESIAN Optimization	0.82
[7]	2020	Cross-validation + Forward feature selection method	Random Forest	0.82
[10]	2021	filter, wrapper, and unsupervised method	Improved Deep Belief Network (IDBN)	0.94
[13]	2022	----------	SVM & XGBoost	0.88 , 0.91
[11]	2022	Ada Boost, SVM, Random Forest, Lasso Regression	Deep Recurrent Neural Networks (DNN)	0.87
[19]	2023	Boruta Algorithm & PCA	Wide & deep Neural Network (DNN)	0.99
[18]	2023	Random forest	SVM	0.89
[24]	2023	PCA Analysis	Deep neural network with six hidden layers (DNN)	0.72

5.3. The Most Common Data Sources and Datasets

The growing application of artificial intelligence (AI) in Alzheimer's disease (AD) research has been greatly supported by the availability of large, high-quality datasets. Several major repositories serve as foundational sources of multimodal data, particularly in clinical, imaging, gene expression, and genomic domains. Among these, the Alzheimer's Disease Neuroimaging Initiative (ADNI), Gene Expression Omnibus (GEO), the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS), and Genome-Wide Association Studies (GWAS) represent critical resources for training and validating machine learning (ML) and deep learning (DL) models. Each resource provides a unique data type that supports specific AI-driven tasks such as disease classification, progression modeling, biomarker discovery, and risk prediction.

5.3.1. Alzheimer's Disease Neuroimaging Initiative (ADNI)

ADNI is one of the most comprehensive and widely used datasets in AD research. It includes longitudinal clinical assessments, structural and functional neuroimaging (MRI and PET), cerebrospinal fluid (CSF) biomarkers, and genetic information (including APOE genotypes). The dataset is particularly well-suited for training AI models due to its multimodal nature and standardized data collection.

In AI-based studies, ADNI is commonly used to develop models for predicting conversion from mild cognitive impairment (MCI) to AD, classifying disease stages, and modeling neurodegeneration patterns. Convolutional neural networks (CNNs) have been applied to MRI and PET images to detect structural brain

atrophy and amyloid accumulation, while traditional ML models have utilized clinical and biomarker data for diagnostic classification and cognitive decline prediction.

5.3.2. Gene Expression Omnibus (GEO)

GEO is a public repository that hosts a broad range of transcriptomic datasets, including microarray and RNA sequencing (RNA-Seq) data from both brain tissue and peripheral samples. These datasets capture gene expression levels, providing molecular insights into the mechanisms underlying AD pathology.

In AI research, gene expression datasets from GEO are used to build classifiers that distinguish AD from control samples based on molecular profiles. Researchers apply dimensionality reduction techniques (e.g., PCA, autoencoders), followed by supervised learning algorithms (e.g., support vector machines, random forests, deep neural networks) to identify gene signatures associated with disease state. These datasets are also valuable for unsupervised clustering to uncover molecular subtypes and for feature selection in biomarker discovery pipelines.

5.3.3. National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS)

NIAGADS serves as a centralized platform for accessing genomic datasets related to AD, including whole-genome sequencing (WGS), whole-exome sequencing (WES), single nucleotide polymorphism (SNP) genotyping, and GWAS summary statistics. It integrates data from multiple cohorts and studies, enabling comprehensive analyses of genetic contributions to AD.

AI applications leveraging NIAGADS data often focus on genetic risk prediction and variant prioritization. Machine learning models are trained to predict disease susceptibility using polygenic risk scores or selected SNP features. In some studies, multi-omics integration is employed to combine genomic data from NIAGADS with expression or proteomic data, enhancing model performance and interpretability.

5.3.4. Genome-Wide Association Studies (GWAS)

GWAS are large-scale studies designed to identify common genetic variants associated with disease traits, including AD. While GWAS results are often included in databases like NIAGADS, they are also published independently and form the basis for many AI-enabled analyses.

In AI-based AD research, GWAS results are primarily used for feature engineering and gene prioritization. SNPs identified as significant are used to construct predictive models or to inform gene-level analyses. Some researchers apply deep learning models to GWAS data to capture non-linear interactions between variants or integrate GWAS findings with gene expression data to identify functional targets and regulatory mechanisms.

Table 4. Summary of dataset studies and their usage for AD detection

Repository/ Study	Main Data Types	AI Applications
ADNI	Clinical, Imaging, Genotype	Disease classification, progression modeling, imaging-based diagnosis
GEO	Gene Expression (RNA-Seq, Microarray)	Gene-based classification, biomarker discovery, clustering
NIAGADS	Genomic (SNPs, GWAS, WGS/WES)	Genetic risk modeling, variant prioritization, feature selection
GWAS	SNP Association Data	Polygenic modeling, gene prioritization, genetic risk prediction

5.4. Challenges of Machine Learning Approaches in Alzheimer's Disease Detection

Machine learning methods have demonstrated potential in detecting Alzheimer's disease but still there are some limitations and challenges that researchers face. Some of the shortages or areas where machine learning approaches in the detection of Alzheimer's disease can be improved are: Limited and imbalanced datasets: One of the key challenges in developing robust machine learning models for Alzheimer's disease detection is the availability of high-quality and diverse datasets. Obtaining large and well-balanced datasets, including data from different demographics and stages of the disease, can be challenging. This scarcity of data can limit the generalizability and performance of machine learning models. Interpretability and explain ability: Machine learning models, especially those leveraging deep learning methods, often face challenges with interpretability. Their black-box nature complicates the understanding of the features or biomarkers driving the predictions.

Interpretable machine learning methods that offer insights into decision-making are crucial for fostering trust and acceptance in clinical settings. Early detection and prediction**:** Early detection of Alzheimer's disease is crucial for effective intervention and treatment. However, most machine learning approaches focus on diagnosing the disease at later stages when symptoms are more pronounced. Developing accurate and reliable models for early detection and prediction of Alzheimer's disease remains a challenge, as subtle signs and biomarkers at early stages are not well understood. Lack of standardized protocols for data acquisition, preprocessing, feature selection, and evaluation metrics in Alzheimer's disease research. This leads to inconsistencies in the reported results and makes it difficult to compare and reproduce different machine learning approaches. Establishing standardized benchmarks and evaluation protocols would facilitate better comparison and validation of algorithms.

Given the growing use of machine learning in healthcare, ethical considerations and privacy concerns are of utmost importance. Protecting the privacy and security of sensitive patient data during the training of machine learning models is critical. Furthermore, it is essential to address potential biases in both the data and algorithms to prevent disparities in diagnosis and treatment outcomes.

Addressing these shortages and challenges requires collaboration between researchers, clinicians, and data scientists. It involves the collection of large, diverse, and well-annotated datasets, development of interpretable and explainable machine learning models, focus on early detection and prediction,

establishment of standardized evaluation protocols, and adherence to ethical guidelines and privacy regulations.

6. Conclusion

This review provides a comprehensive and systematic overview of the existing literature on Alzheimer detection research. It summarizes the current state of research and highlights key findings, methodologies, and advancements in the field of using genetic data and AI for AD detection. The overview helps researchers and clinicians stay updated with the latest developments and understand the overall landscape of the field. Moreover, the review critically evaluates the various methods and approaches employed in utilizing genetic data and AI for AD detection. It assesses the strengths, limitations, and performance of different AI techniques in predicting AD risk or diagnosing AD, such as machine learning algorithms, deep learning architectures, or polygenic risk score models. This evaluation guides researchers in selecting appropriate methods for their own studies and identifies areas that require further improvement or investigation.

The review explores the integration of genetic data with other data types and highlights the potential synergies and challenges associated with these integrative approaches. This integration of multiple data types can lead to more accurate and reliable AI models for AD detection.

References

[1]. 2023 Alzheimer's disease facts and figures. Alzheimer's & dementia : the journal of the Alzheimer's Association. (2023). 19(4), 1598–1695. https://doi.org/10.1002/alz.13016.
[2]. Freudenberg-Hua, Y., Li, W., & Davies, P. The Role of Genetics in Advancing Precision Medicine for Alzheimer's Disease-A Narrative Review. Frontiers in medicine. (2018). 5, 108. https://doi.org/10.3389/fmed.2018.00108.
[3]. Kurtakoti, A.U., Hiremath, N.D., Patil, N.S., & Rane, A. Benefits of Early Detection of Alzheimer's Disease—A Machine Learning with Image Processing Approach. Journal of Computational and Theoretical Nanoscience. (2020). 17, 378-383.
[4]. S. Perera, K. Hewage, C. Gunarathne, R. Navarathna, D. Herath and R. G. Ragel. Detection of Novel Biomarker Genes of Alzheimer's Disease Using Gene Expression Data. Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka. (2020). pp. 1-6. https://doi.org/10.1109/MERCon50084.2020.9185336.
[5]. H. Ahmed, H. Soliman and M. Elmogy. Early Detection of Alzheimer's Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques.International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain. 2020. pp. 1-6.. https://doi.org/10.1109/ICDABI51230.2020.9325640.
[6]. Park, C., Ha, J., & Park, S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Syst. (2020). Appl., 140.
[7]. Rangaswamy, U., Dharshini, S. A. P., Yesudhas, D., & Gromiha, M. M. VEPAD - Predicting the effect of variants associated with Alzheimer's disease using machine learning. Computers in biology and medicine. 2020. 124, 103933. https://doi.org/10.1016/j.compbiomed.2020.103933.

[8] S. Khanal, J. Chen, N. Jacobs and A. -L. Lin. Alzheimer's Disease Classification Using Genetic Data. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021. pp. 2245-2252. https://doi.org/10.1109/BIBM52615.2021.9669730.
[9]. Mohamed, M.O., Salem, N.M., & Ghoneim, V.F. Exploring the Efficiency of Hub Genes in Identification of Alzheimer Disease. 38th National Radio Science Conference (NRSC). 2021. 1, 243-250.
[10]. Mahendran, N., Vincent, P. M. D. R., Srinivasan, K., & Chang, C. Y. Improving the Classification of Alzheimer's Disease Using Hybrid Gene Selection Pipeline and Deep Learning. Frontiers in genetics. 2021. 12, 784814. https://doi.org/10.3389/fgene.2021.784814
[11]. Mahendran, N., & P M, D. R. V. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. Computers in biology and medicine. (2022). 141, 105056. https://doi.org/10.1016/j.compbiomed.2021.105056
[12]. Jo, T., Nho, K., Bice, P., Saykin, A. J., & Alzheimer's Disease Neuroimaging Initiative. Deep learning-based identification of genetic variants: application to Alzheimer's disease classification. Briefings in bioinformatics. 2022. 23(2), bbac022. https://doi.org/10.1093/bib/bbac022
[13]. Abbas, Z., Tayara, H., & Chong, K.T.. Alzheimer's disease prediction based on continuous feature representation using multi-omics data integration. Chemometrics and Intelligent Laboratory Systems. 2022.
[14]. Ahmed, H., Soliman, H., & Elmogy, M.. Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree. Computers in biology and medicine. 2022. 146, 105622. https://doi.org/10.1016/j.compbiomed.2022.105622
[15]. Bae, J., Logan, P. E., Acri, D. J., Bharthur, A., Nho, K., Saykin, A. J., Risacher, S. L., Nudelman, K., Polsinelli, A. J., Pentchev, V., Kim, J., Hammers, D. B., Apostolova, L. G., & Alzheimer's Disease Neuroimaging Initiative. A simulative deep learning model of SNP interactions on chromosome 19 for predicting Alzheimer's disease risk and rates of disease progression. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2023;19(12), 5690–5699. https://doi.org/10.1002/alz.13319
[16]. Hermes, S., Cady, J., Armentrout, S., O'Connor, J., Carlson, S., Cruchaga, C., Wingo, T., Greytak, E. M., & Alzheimer's Disease Neuroimaging Initiative. Epistatic Features and Machine Learning Improve Alzheimer's Risk Prediction Over Polygenic Risk Scores. medRxiv : the preprint server for health sciences, 2023.02.10.23285766. https://doi.org/10.1101/2023.02.10.23285766.
[17]. Brookes KJ, Guetta-Baranes T, Thomas A and Morgan K.. An alternative method of SNPinclusion to develop a generalized polygenic risk score analysis across Alzheimer's disease cohorts. Front. Dement. 2023. 2:1120206.doi: 10.3389/frdem.2023.1120206.
[18]. Alatrany, A. S., Khan, W., Hussain, A. J., Mustafina, J., & Al-Jumeily, D.. Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data. IEEE/ACM transactions on computational biology and bioinformatics. 2023; 20(5), 2700–2711. https://doi.org/10.1109/TCBB.2022.3233869
[19]. Alatrany AS, Khan W, Hussain A, Al-Jumeily D, for the Alzheimer's Disease Neuroimaging Initiative. Wide and deep learning based approaches for classification of Alzheimer's disease using genome-wide association studies. PLoS ONE. 2023. 18(5): e0283712.https://doi.org/10.1371/journal.pone.0283712.
[20]. Alshamlan, H., Omar, S., Aljurayyad, R., & Alabduljabbar, R.. Identifying Effective Feature Selection Methods for Alzheimer's Disease Biomarker Gene Detection Using Machine Learning. Diagnostics (Basel, Switzerland). 2023. 13(10), 1771. https://doi.org/10.3390/diagnostics13101771.

[21]. A Y. Ay, Y. H. Choo, A. Bhatti and C. P. Lim. Classification of Inflammatory Gene Expression Patterns with Machine Learning Models. IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML), Urumqi, China, 2023. pp. 115-119. https://doi.org/10.1109/PRML59573.2023.10348265.
[22]. W. DeGroat, V. Venkat, W. Pierre-Louis, H. Abdelhalim, and Z. Ahmed. Hygieia: Ai/ml pipeline integrating healthcare and genomics data to investigate genes associated with targeted disorders and predict disease. Softw. Impacts. 2023. vol. 16, p. 100493. https://doi.org/10.1016/j.simpa.2023.100493.
[23]. Abd El Hamid, M.M., Shaheen, M., Omar, Y.M., & Mabrouk, M.S.. Discovering epistasis interactions in Alzheimer's disease using integrated framework of ensemble learning and multifactor dimensionality reduction (MDR). Ain Shams Engineering Journal. 2023.
[24]. Shigemizu, D., Akiyama, S., Suganuma, M., Furutani, M., Yamakawa, A., Nakano, Y., Ozaki, K., & Niida, S.. Classification and deep-learning-based prediction of Alzheimer disease subtypes by using genomic data. Translational psychiatry. 2023. 13(1), 232. https://doi.org/10.1038/s41398-023-02531-1
[25]. Li, J., Chen, D., Liu, H., Xi, Y., Luo, H., Wei, Y., Liu, J., Liang, H., & Zhang, Q.. Identifying potential genetic epistasis implicated in Alzheimer's disease via detection of SNP-SNP interaction on quantitative trait CSF Aβ42. Neurobiology of Aging. 2023. 134, 84-93.
[26]. Finney, C. A., Delerue, F., Gold, W. A., Brown, D. A., & Shvetcov, A.. Artificial intelligence-driven meta-analysis of brain gene expression identifies novel gene candidates and a role for mitochondria in Alzheimer's disease. Computational and structural biotechnology journal. 2023. 21, 388–400. https://doi.org/10.1016/j.csbj.2022.12.018
[27]. Rebelo, A.A. Unsupervised Learning of Physical Models: Uses and Limitations of Principal Component Analysis. 2017.